CGIAR Research Program on
Climate Change, Agriculture and Food Security (CCAFS)

Transition from Raw to Primary Data
Video Transcript

October 2013

These materials were produced for and with funding from the Climate Change Agriculture and Food Security Research Program of the Consultative Group on International Agricultural Research (CGIAR).

Introduction

Once the data has been collected for your activity it needs to follow a process from being in its original raw state to transition through to the final primary datasets that will be analysed.

This video introduces the key stages in this transition process.

Raw Data

There are several formats raw data can take; the most common are the results from paper based and electronic data collection activities, this tends to be 'new' data which is collected as part of the activity. Other data from external sources, such as climate data from local a weather station, or historical sources such as baseline survey data, can also be used to create primary data for an activity.

Once the data has been collected, or gathered from other sources, it can enter the transition process from raw to primary data.

The Transition Process

The first stage is planning, and this stage overlaps with work previously conducted when producing data collection tools.

Planning

The planning stage involves thinking about how each piece of data is going to be handled and stored. This stage is required to be able to develop data entry software. Activities usually involve numerous pieces of data and decisions need to be made prior to data entry about: which text data is going to be coded and which is going to be stored as a free text field, which data are numerical, how are missing values going to be entered, which variables are key identifier variables and as such need to be on all relevant datasets etc.

Activities also need to plan if dedicated data entry software such as CS-Pro is going to be used, or if they are going to use a more general software such as Microsoft Excel.

More details about this stage can be found in the 'Storing Numerical and Non Numerical Information' video and corresponding document.

Data Entry

For paper based data this can be achieved using software of varying complexities and abilities. If using specialist software (such as Access or CS-Pro) an entry system should be designed to incorporate checks whilst ensuring it remains intuitive for the staff to use. We recommend double data entry, to reduce errors in the data due to this stage in the process.

If using electronic data collection devices the data collection and entry are combined into one step; think ahead and provide spare parts/additional devices and regularly back-up the data to avoid problems! Electronic data collection does mean data entry errors cannot be eliminated through a

process like double data entry, so ensure the data collection staff are thoroughly trained as the activity will be reliant on the original data entered.

Data Manipulation

This is the creation of new or derived variables and datasets.

Analysis usually involves derived variables; these are variables created, or derived, from those in the raw data. For example the age of the participants in groups rather than the raw exact ages, or yields in kg per hectare from the raw yield and plot size measurements.

Similarly entire derived datasets can be created, for example taking daily climate data and summarising it so that a new derived dataset has the monthly averages which are then used in the analysis, or combining the external or historical data with that collected by the activity.

It is important to keep documentation detailing the creation of the derived variables and datasets; you should keep the program code or syntax used to perform the manipulations as part of the activity Metadata.

Throughout the whole transition process data should be constantly checked for quality and validated. Data checks prior to the data manipulation stage are purely about checking the individual pieces of data, derived variables and derived datasets should also be rigorously checked – at this stage you are no longer focusing on data issues, but checking the manipulations have been conducted correctly.

Primary Data

Once the raw data has been through the transition process of planning, data entry and data manipulation the result is the primary data. Primary datasets should contain all of the variables required to conduct the analysis outlined in the analysis plan.

More details about the Transition from Raw to Primary data introduced in this video are available in the corresponding guide as part of the Data Management Support Pack.

Appendix I – CCAFS Data Management Support Pack

This document is part of the CCAFS Data Management Support Pack produced by the Statistical Services Centre, University of Reading, UK. The following materials are available in the pack:

0. Data Management Strategy
    a. CCAFS Data Management Strategy
1. Research Protocols
    a. Writing Research Protocols – a statistical perspective
    b. Preparation of Research Protocols – Good Practice Case Study
    c. What is a Research Protocol, and how to use one (Video & Transcript)
    d. Details of what a Research Protocol should contain (Video & Transcript)
2. Data Management Policies & Plans
    a. Creating a Data Management Plan
    b. Data Management Plan (Video & Transcript)
    c. Example Data Management Activity Plan
    d. Example Consent Form
3. Budgeting & Planning
    a. Budgeting & Planning for Data Management
    b. ToR Data Support Staff
    c. Budgeting & Planning (Video & Transcript)
4. Data Ownership
    a. Data Ownership and Authorship
    b. Template – Data Ownership Agreement
    c. CCAFS Data Ownership & Sharing Agreement
    d. Data Ownership & Authorship (Video & Transcript)
5. Data & Document Storage
    a. Creating and Using a DDS
    b. DDS Introduction – (Video & Transcript)
    c. DDS Organisation – (Video & Transcript)
    d. DDS Ownership – (Video & Transcript)
    e. Introduction to Dropbox – (Video & Transcript)
6. Archiving & Sharing
    a. Archiving & Sharing Data
    b. Data and Documents to Submit for Archiving – a checklist
    c. MetaData
    d. Archiving & Sharing (Video & Transcript)
    e. Metadata (Video & Transcript)
    f. CCAFS HBS Questionnaire
    g. CCAFS HHS Code Book
    h. CCAFS Training Manual for Field Supervisors

7. CCAFS Data Portals
    a. Portals for CCAFS Outputs
    b. AgTrials Summary
    c. CCAFS-Climate Summary
    d. DSpace Introduction
    e. Introduction to Dataverse (Video & Transcript)
    f. Creating a Dataverse (Video & Transcript)
    g. Dataverse Study Catalogue
    h. CCAFS Dataverse (Video & Transcript)
8. Data Quality & Organisation
    a. Data Quality Assurance
    b. Guidance for handling different types of Data
    c. Transition from Raw to Primary Data
    d. Data Quality Assurance (Video & Transcript)
    e. Guidance for handling different types of data (Video & Transcript)
    f. Transition from Raw to Primary Data (Video & Transcript)